A nonparametric change point model for multivariate phase-II statistical process control

Mark Holland
Douglas Hawkins

School of Statistics
University of Minnesota

May 24, 2011

Statistical Process Control (SPC) definitions

Statistical Process Control refers to a collection of tools designed to detect a shift in distribution of a sequence of observations.

- **Phase-I SPC**: Analysis is performed on a fixed set of historical data.
- **Phase-II SPC**: Ongoing analysis is performed on a possibly never-ending stream of observations.

- **Common cause variability** is inherent variability in a process, even when running as designed.
- **Special cause variability** is not a normal part of the process, but is the result of the intrusion of an unexpected factor.

A process is **in control** when only common cause variability exists, but is **out of control** when special cause variability is introduced.

Statistical Process Control (SPC) applications

Traditionally used in manufacturing settings, but developments in modern industries have created demand for new monitoring techniques

- Health care (Thor et al. 2007)
  - Laboratory setting, e.g. Chemical assay methods
  - Direct patient care, e.g. ICU vital signs
- Post-market product performance
- Groundwater and air quality

Many current applications require multivariate nonparametric methods

- Several measurements must be monitored simultaneously
- Multivariate normal distribution rarely applies
- Difficult to check if a data set follows multivariate normal distribution

Aluminum Smelter Data

- Aluminum smelting refers to an electrolysis process to reduce refined aluminum ore into metallic aluminum.
- Data set consists of alumina (Al_2O_3) content of a smelter feed along with several impurities: silica (SiO_2), ferric oxide (Fe_2O_3), magnesium oxide (MgO), and calcium oxide (CaO).
- As expected with compositional data, content of compounds are negatively correlated.
- Monitor for change in composition of alumina or any of the impurities.

Standard SPC tools

Some traditional phase-II SPC methods include

- Shewart Chart, Cumulative sum (CUSUM), Exponentially weighted moving average (EWMA)

Limitations of traditional methods

- In-control distribution including all parameters must be known.
- In practice, parameter estimates from a phase-I training sample are typically substituted for the truth.
- In some applications a large historical training sample is not available, so monitoring must begin shortly after data collection begins.
  - ICU vital signs
  - pollution control monitoring
- Must be "tuned" to detect a specific size of shift.

Change point approach to phase-II SPC

- Hawkins, Qiu, and Kang (2003) proposed change point model for phase-II SPC, which does not require knowledge of in- or out-of-control process parameters.
- Skeleton of change point approach:
  1. Choose two-sample test statistic for comparing left- and right-segments of process readings, $\{X_1, \ldots, X_k\}$ and $\{X_{k+1}, \ldots, X_n\}$.
  2. Apply test for all possible split-points, $k = 1, 2, \ldots, n-1$.
  3. If maximum test statistic value is outside of control limits, signal that a shift has occurred. Otherwise, collect another observation and repeat.
- Originally implemented with likelihood ratio test for shift in mean for univariate normal data
- Zamba and Hawkins (2006) extended using likelihood ratio test for shift in multivariate normal data
- Deng (2009) extended using univariate Wilcoxon-Mann-Whitney nonparametric test for difference in location

Rank based multivariate change point model

We used existing hypothesis test proposed by Choi and Marden (1997) to design a change point model for phase-II SPC use.

- We observe n random vectors from a multivariate location family distribution

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k \sim F(\boldsymbol{\mu})$$
$$\mathbf{X}_{k+1}, \mathbf{X}_{k+2}, \ldots, \mathbf{X}_n \sim F(\boldsymbol{\mu} + \boldsymbol{\delta}).$$

  and we wish to test

$$H_0 : \boldsymbol{\delta} = \mathbf{0} \; vs.$$
$$H_a : \boldsymbol{\delta} \neq \mathbf{0}$$

Multivariate nonparametric test (Choi and Marden 1997)

Suppose we observe a sample of $p \times 1$ random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$. For $1 \leq i, j \leq n$, define

$$\mathbf{D}_{ij} = \frac{\mathbf{X}_i - \mathbf{X}_j}{||\mathbf{X}_i - \mathbf{X}_j||}$$

and for $1 \leq i \leq n$, define

$$\mathbf{R}_n(\mathbf{X}_i) = \sum_{j=1}^{n} \mathbf{D}_{ij}.$$

Then, $\mathbf{R}_n(\mathbf{X}_i)$ is the centered directional rank vector of $\mathbf{X}_i$.

Multivariate nonparametric test *(cont'd)*

Next, let

$$\bar{\mathbf{R}}_n^{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathbf{R}_n(\mathbf{X}_i).$$

and define the covariance matrix estimator

$$\hat{\Sigma}_{R_{k,n}} = \frac{n-k}{(n-1)nk}\sum_{i=1}^{n} \mathbf{R}_n(\mathbf{X}_i)\mathbf{R}_n(\mathbf{X}_i)'.$$

Finally, define the test statistic

$$R_{k,n} = \bar{\mathbf{R}}_n^{(k)\prime}\hat{\Sigma}_{R_{k,n}}^{-1}\bar{\mathbf{R}}_n^{(k)}.$$

Under mild conditions, $R_{k,n}$ has asymptotic null distribution χ_p^2.

Multivariate nonparametric change point model

- Test statistic for existence of a change point

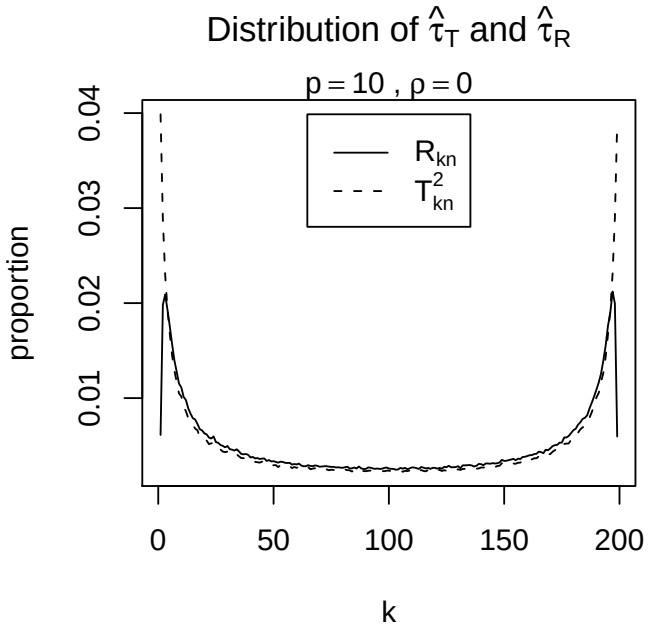
$$R_{\max,n} = \max_{1 \leq k \leq n-1} R_{k,n}$$

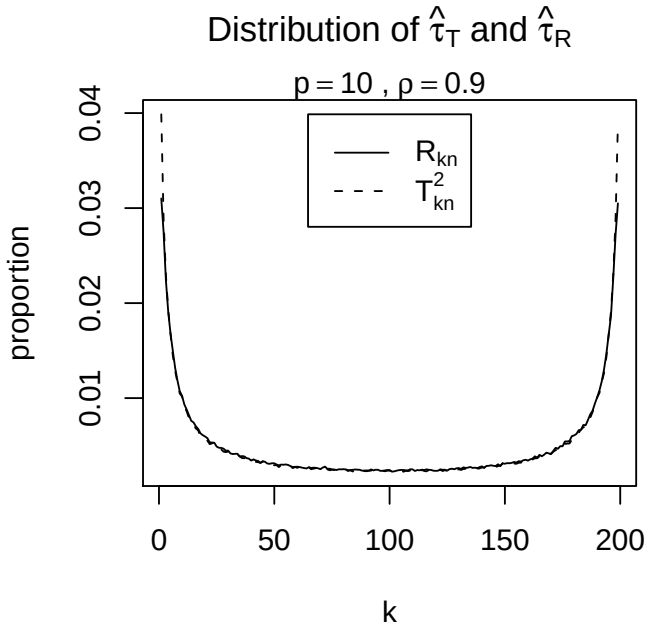
- Estimate of the location of the change point

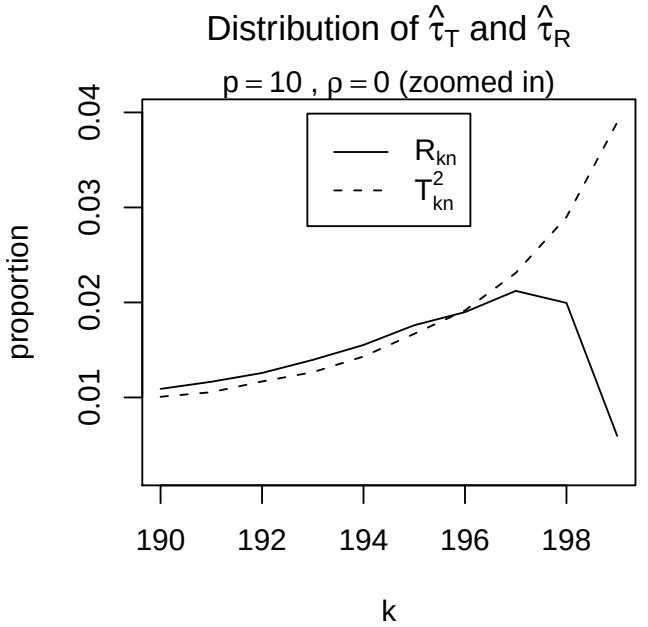
$$\hat{\tau}_{R,n} = \arg \max_{1 \leq k \leq n-1} R_{k,n}$$

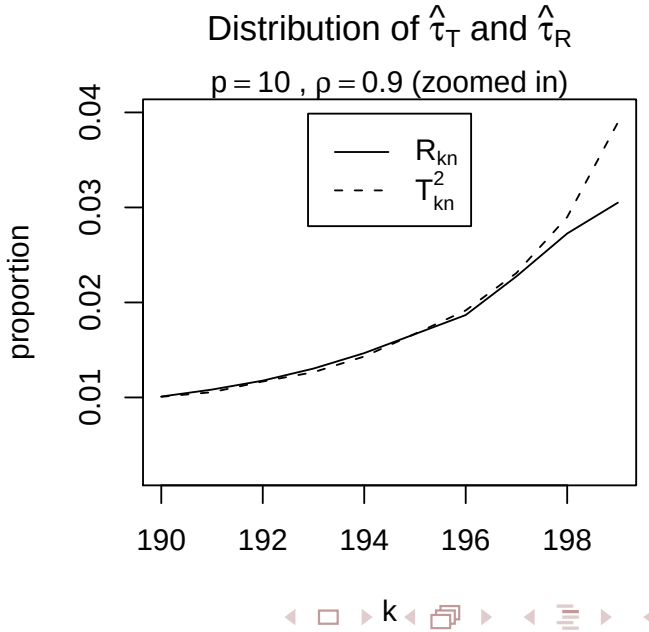
Fixed-sample size simulation results

- When both k and $n-k$ are large, the distribution of $R_{k,n}$ is approximately χ^2_p, as expected $(k = 100, n-k = 50)$.
- When k or $n-k$ is small, the distributions of $R_{k,n}$, $R_{\max,n}$, and $\hat{\tau}_{R,n}$ are affected by the dependence structure of the simulated data.
- The following plots show the estimated distribution of the location of the maximum $R_{k,n}$ value for a sample of $n = 200$ equicorrelated MVN random vectors with $\rho = 0, 0.9$.


Distribution of $\hat{\tau}_T$ and $\hat{\tau}_R$
p = 10 , ρ = 0
proportion
0.01
0.02
0.03
0.04
R_{kn}
T^2_{kn}
0
50
100
150
200
k


Distribution of $\hat{\tau}_T$ and $\hat{\tau}_R$
p = 10 , ρ = 0.9
proportion
0.01
0.02
0.03
0.04
R_{kn}
T^2_{kn}
0
50
100
150
200
k


Distribution of $\hat{\tau}_T$ and $\hat{\tau}_R$
p = 10 , ρ = 0 (zoomed in)
proportion
0.01
0.02
0.03
0.04
R_{kn}
T^2_{kn}
190
192
194
196
198
k


Distribution of $\hat{\tau}_T$ and $\hat{\tau}_R$
p = 10 , ρ = 0.9 (zoomed in)
proportion
0.01
0.02
0.03
0.04
R_{kn}
T^2_{kn}
190
192
194
196
198
k

Quarantine

- Problem: Distribution of $\hat{\tau}$ depends on dependence structure of data
  - Distribution only depends on dependence structure when split point is near the boundary of the sequence of data
- Solution: **Quarantine**, that is restrict search for a change point to interior of sequence.

Quarantined Phase-II SPC procedure

To use $R_{k,n}$ for phase-II SPC:

- Collect observation $\mathbf{X}_n$ and compute

$$R_{\max,n,c} = \max_{c<k<n-c} R_{k,n}$$

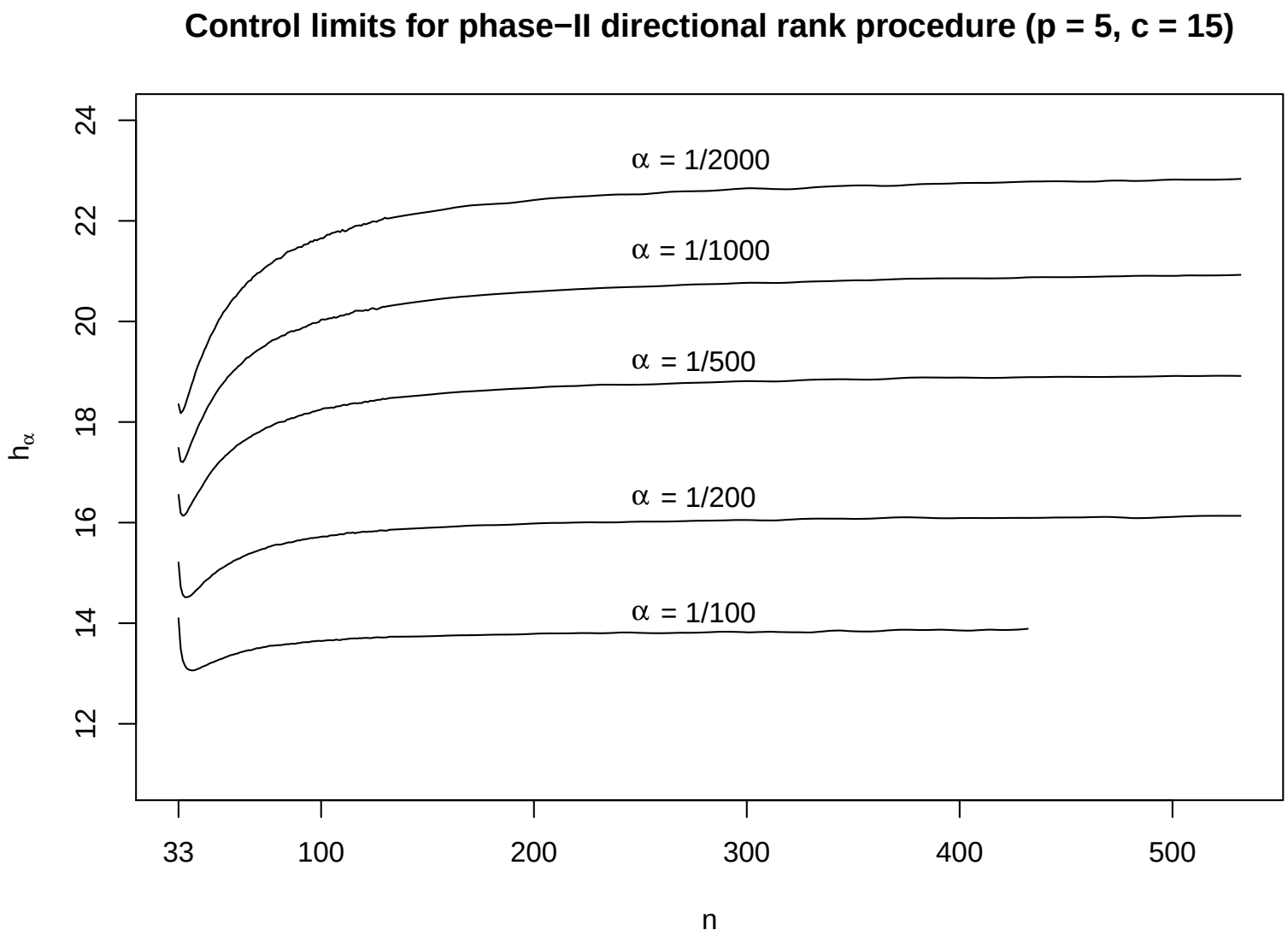
- If $R_{\max,n,c} > h_{n,\alpha,p,c}$ signal that a shift has occurred, otherwise collect another observation and repeat.
- c determines the number of observations at each end of the sequence that are quarantined.

Choose control limits such that probability of false signal is constant across time. When process is in control,

$$P[R_{\max,n,c} > h_{n,\alpha,p,c} | R_{\max,j,\alpha,c} \leq h_{j,\alpha,p,c}; j < n] = \alpha.$$

Use Monte Carlo simulation to obtain sequence of control limits, $\{h_{n,\alpha,p,c}\}$.

Control limits



Average run length (ARL) as a performance metric

The **average run length (ARL)** of a phase-II SPC procedure is the average number of observations collected before the first signal occurs.

- Design phase-II SPC procedure to control in control ARL to a minimum value, $1/\alpha$.
- Subject to constraint on in control (IC) ARL, we would like to minimize out of control (OOC) ARL.
- Similar to common goal in hypothesis testing
  - Minimize Type-II error rate given that Type-I error rate is controlled to level α.

In control ARL simulation results

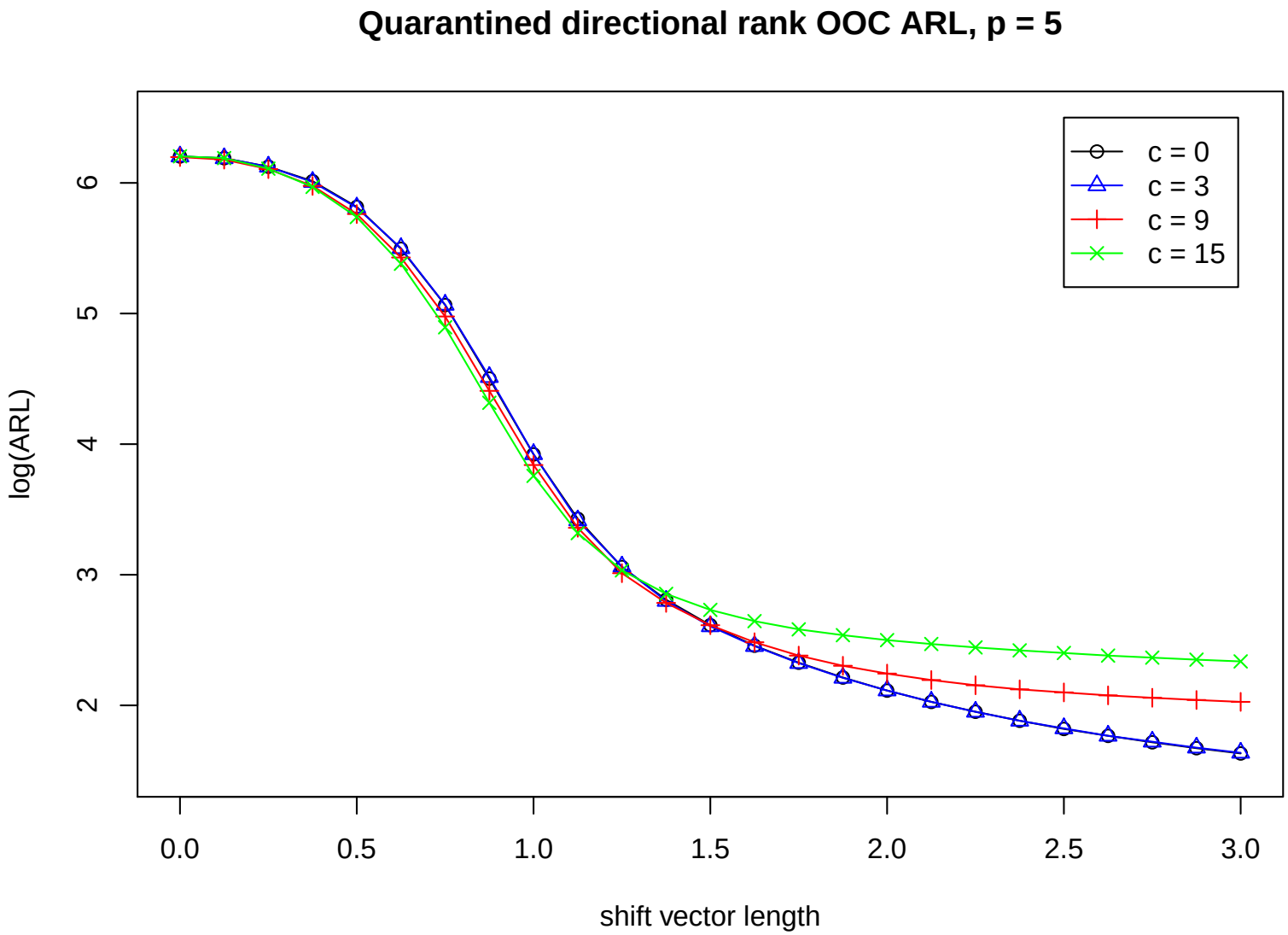
Simulated equicorrelated data with correlation $\rho = 0, 0.5, 0.9$
Default quarantine values: $c = 9$ for $p = 2$; $c = 15$ for $p = 5, 10$

- Multivariate Normal Data:
  - Default quarantine is sufficient to achieve IC ARL within 10% of nominal for all values of p and ρ considered
- Multivariate Gamma Data:
  - Positive, right-skewed distribution. Not elliptically symmetric.
  - Default quarantine is sufficient to achieve IC ARL within 10% of nominal, except when $p = 5, 10$ and $\rho = 0.9$
- Multivariate Cauchy Data:
  - Symmetric distribution, much heavier tails than MVN distribution.
  - Default quarantine is sufficient to achieve IC ARL within 10% of nominal, except when $p = 10$ and $\rho = 0.5, 0.9$

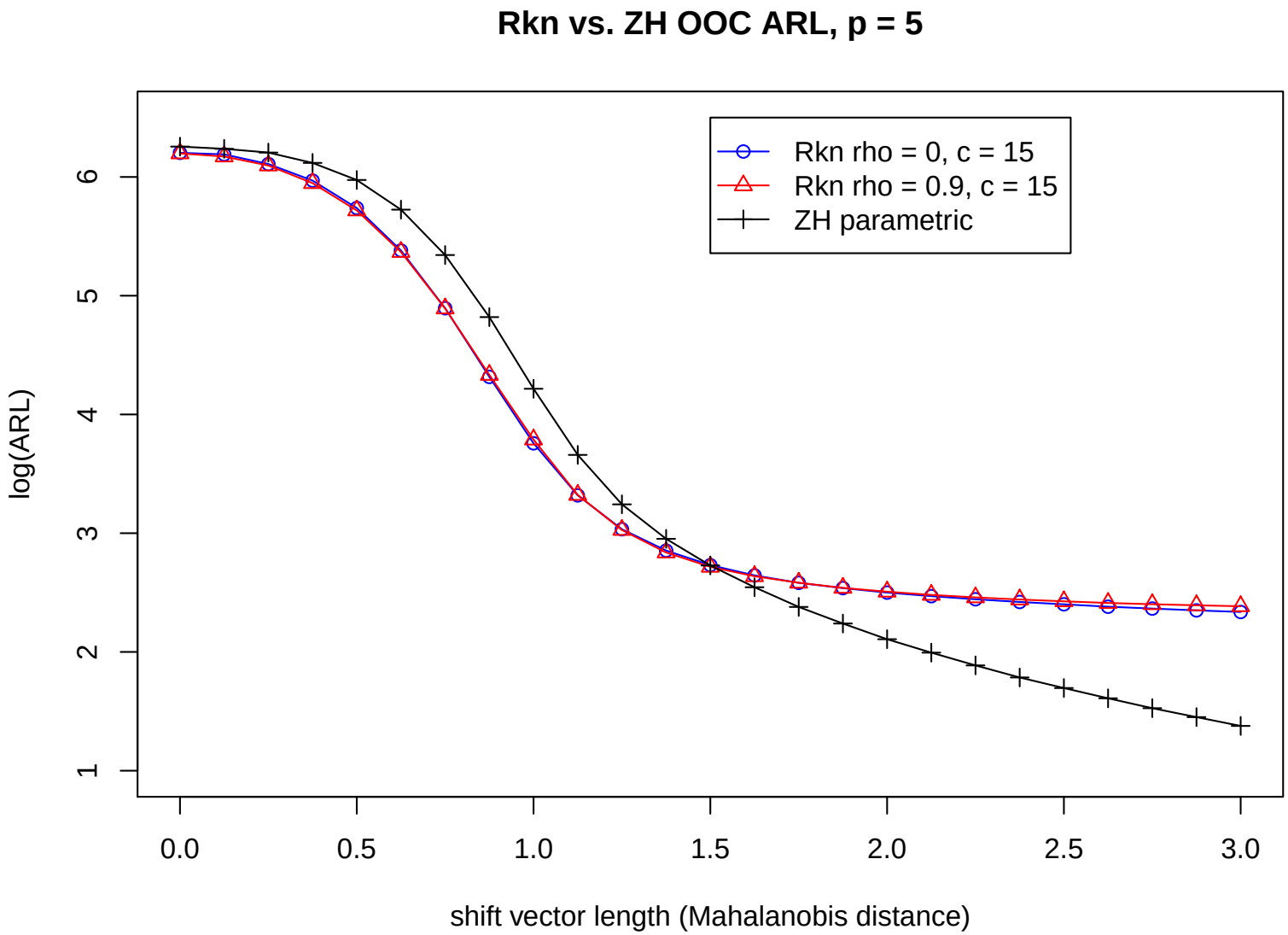
Out of control simulation methodology

1. Simulate $n = 32$ equicorrelated in control observations from the multivariate normal distribution with $p = 5$ and mean vector $\boldsymbol{\mu} = \mathbf{0}$.
2. Introduce mean vector shift $\boldsymbol{\delta} = (\delta, \ldots, \delta)^T$ and begin monitoring with quarantine $c = 15$ at observation $n = 33$.
3. Simulate data sequence until signal occurs using control limits chosen to achieve in control ARL $1/\alpha = 500$.
4. record run length $=$ number of observations collected since monitoring began.
5. Repeat for 100,000 simulated data sequences and compute ARL.

Effect of quarantine on out of control ARL



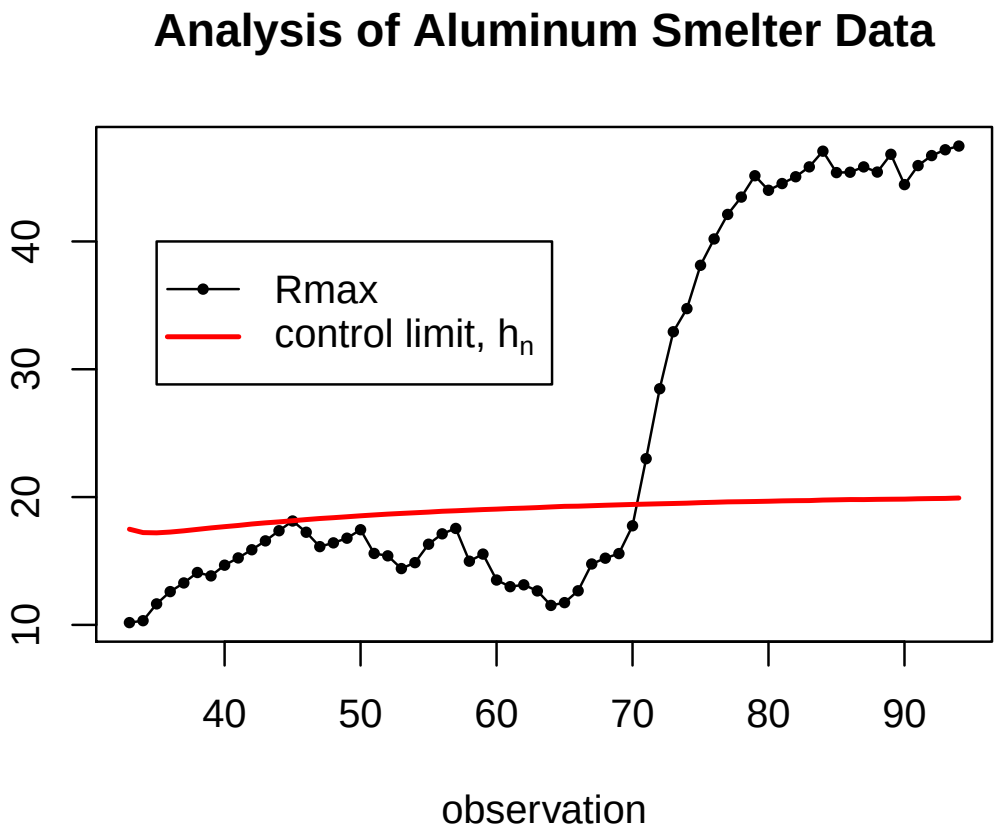
Performance comparison with parametric method



Diagnostic to select degree of quarantine

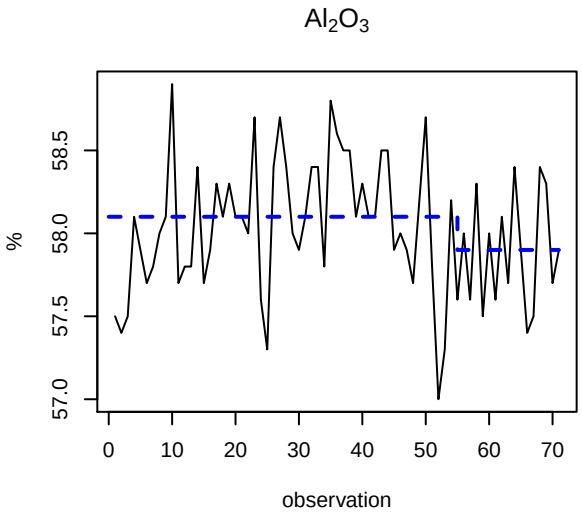
- Based on **copula** function: A **copula** is a p-dimensional distribution function on $[0,1]^p$ with uniform univariate marginal distributions.
- **Sklar's theorem:** any p-dimensional distribution function is associated with a unique copula function.
- Copula can therefore be used to characterize the dependence between the components of a random vector.
- Diagnostic based on Anderson-Darling test for Goodness-of-Fit of multivariate normal copula.

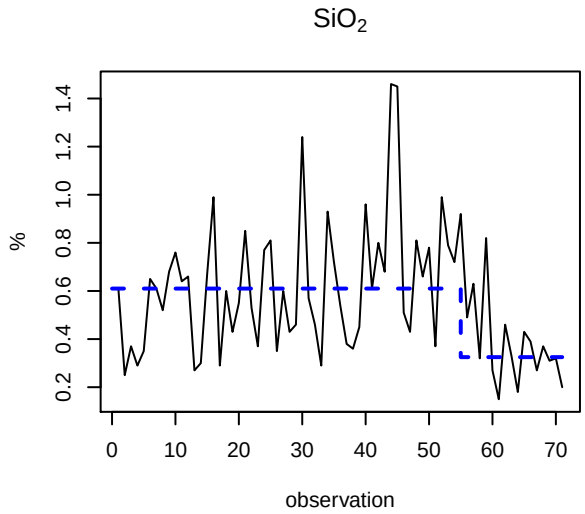
Analysis of Aluminum Smelter Data

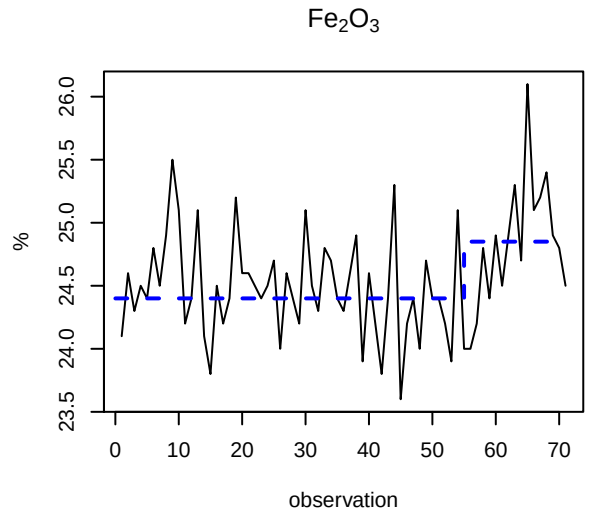


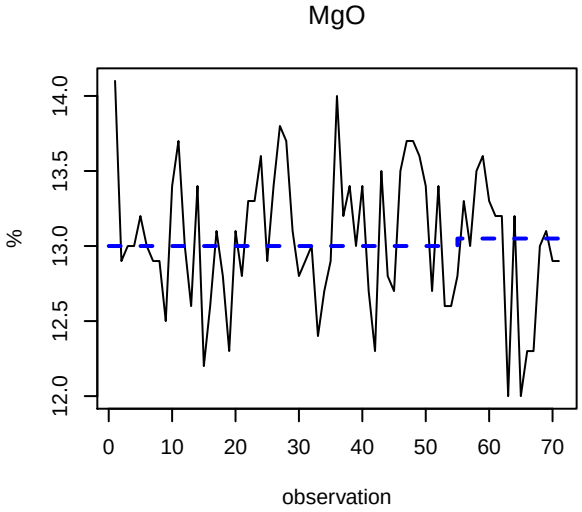
- In control ARL: $1/\alpha = 500$
- Control limit exceeded at observation $n = 71$
- Estimated shift location $\hat{\tau}_{R,n} = 55$

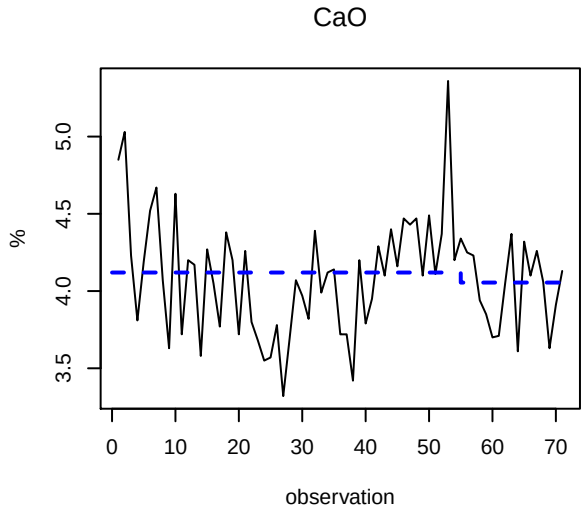
Analysis of Aluminum Smelter Data











Summary

- Traditional SPC methods are not suitable for some modern applications
- Change point model for phase-II SPC does not require phase-I training sample
- Nonparametric multivariate change point model:
  - Does not require assumption of multivariate normality
  - Outperforms parametric method for small to moderate shift sizes, even when data follows multivariate normal distribution
  - Detects large shifts slower than parametric method

References

- Choi, K. and Marden, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance. Journal of the American Statistical Association 92(440), pp. 1581 - 1590.
- Deng, Q. (2009). A nonparametric change-point model for phase II analysis. PhD thesis. University of Minnesota.
- Hawkins, D. M., Qiu, P., and Kang, C. W. (2003). The Changepoint Model for Statistical Process Control. Journal of Quality Technology 35(4), pp. 355-366.
- Thor, J., Lundberg, J., Ask, J. Olsson, J. Carli, C., Harenstam, K., Brommels, M. (2007). Application of statistical process control in healthcare improvement: systematic review. Quality and Safety in Health Care 16, pp. 387-399.
- Zamba, K. D. and Hawkins, D. M. (2006). A multivariate change-point model for statistical process control. Technometrics 48(4), pp. 539-549.

Assumptions required for asymptotic result for Choi and Marden (1997) test statistic:

- Under the Null Hypothesis,

$$\mathbf{\Lambda} = \text{cov}(\mathbf{D}_{ij}) \text{ and } \mathbf{\Omega} = \text{cov}(\mathbf{D}_{ij}, \mathbf{D}_{il})$$

  are finite and positive definite when i, j, and l are all distinct.
- $k/n \longrightarrow \lambda_0 \in (0, 1)$

Multivariate gamma distribution

- Let $Y_0, Y_1, \ldots, Y_p$ be independent gamma random variables with pdf's
$$p_{Y_i}(y_i) = \frac{1}{\Gamma(\theta_i)} e^{-y_i} y_i^{\theta_i - 1}, \ y_i > 0, \ \theta_i > 0.$$
- Define $\mathbf{X} = (Y_0 + Y_1, Y_0 + Y_2, \ldots, Y_0 + Y_p)^T$.
- Marginal distribution of each X_i is a univariate gamma distribution with shape parameter $\theta_0 + \theta_i$.
- $$\rho_{ij} = \text{corr}(X_i, X_j) = \frac{\theta_0}{\sqrt{(\theta_0 + \theta_i)(\theta_0 + \theta_j)}}.$$

Multivariate Cauchy distribution

- Let $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $w \sim \chi^2_\nu$.
- Define

$$\mathbf{X} = \frac{1}{\sqrt{w/\nu}}\mathbf{Y}.$$

- Then, $\mathbf{X}$ follows the multivariate T distribution.
- If $\nu = 1$, $\mathbf{X}$ follows the multivariate Cauchy distribution.